Journal of Phytology 2011, 3(3): 77–82
An Open Access Journal
ISSN: 2075-6240
Available Online: www.journal-phytology.com

REGULAR ARTICLE

STRUCTURE OF DNA BINDING MYB TRANSCRIPTION FACTOR PROTEIN (SCMYBAS1-3) FROM SUGARCANE – THREADING AND *AB INITIO* MODELLING

Prabu G.R.[1,2,3] and Theertha Prasad D.[1,4]

[1]*Molecular Biology & Genetic Engineering Division, Vasantdada Sugar Institute, Manjari (Bk), Tal. Haveli, Pune – 412307, Maharashtra, India*

[2]*Department of Biotechnology, Shivaji University, Vidyanagar, Kolhapur – 416004, Maharashtra, India*

[3]*Department of Biotechnology, School of Life Sciences, Karpagam University, Pollachi Main Road, Eachanari Post, Coimbatore - 641021, TamilNadu, India*

[4]*Department of Biotechnology, University of Agricultural Sciences, GKVK, Bangaluru – 560065, Karnataka, India*

SUMMARY

The sugarcane (*Saccharum officinarum*) stress-related MYB transcription factor gene, *ScMYBAS1-3* demonstrated induced response to water-deficit and salt stress in our previous study. In order to elucidate its sequence-to-structure-to-function paradigm, the putative three-dimensional structure of ScMYBAS1 was generated using threading assembly refinement (I-TASSER) server. Further, PROCHECK, Verify-3D, PROMOTIF and ProSA programs were used to test the quality of model and the scores were within the recommended intervals. The models shed valuable information necessary for future identification of DNA binding regions and the prediction of co-regulated stress induced genes by docking studies.

Key words: I-TASSER; ScMYBAS1; Sugarcane; Water-deficit

Prabu G.R. and D. Theertha Prasad. Structure of DNA Binding MYB Transcription Factor Protein (ScMYBAS1-3) from Sugarcane – Threading and *Ab Initio* Modelling. J Phytol 3/3 (2011) 77-82

*Corresponding Author, Email: prabugajjeraman@gmail.com, Tel.: +91-422-2611146; Fax: +91-422-2611043

1. Introduction

The functional characterization of transcription factor proteins represents a major challenge for biochemical, medical and computational sciences. To keep pace with current structure determination and sequence data analysis, computational protein structure prediction is a dynamic research field steadily increasing in biotechnology communities (Pieper et al., 2004). There are three computational protein structure prediction methods are available (Murzin, 2001): (1) comparative modeling, (2) fold recognition, and (3) ab initio methods. Several reviews have been published recently on various aspects of structure modeling (Kryshtafovych and Fidelis, 2009).

With the improvement of prediction techniques, there are numerous examples where *in silico* models were used to drive further experimental assays like inferring protein function, protein interaction partners and binding site locations, design or improve novel antibodies or enzymes (Zhang, 2009; Pierri et al., 2010).

Among the several prediction techniques, the iterative threading assembly refinement (I-TASSER) server is an integrated platform for automated protein structure and function prediction based on the sequence-to-structure-to-function paradigm (Roy et al., 2010). Starting from an amino acid sequence, I-TASSER first generates three-dimensional (3D) atomic models from multiple threading alignments and iterative structural assembly simulations. The function of the protein is then inferred by structurally matching the 3D models with other known proteins.

Sugarcane (*Saccharum officinarum* L.) is an important industrial sugar crop of tropical and subtropical regions, accounting for almost two-thirds of the world's sugar production. One of the major limitations in sugarcane production is a shortage of water (Menossi et al., 2008), demanding efforts to investigate the adaptation and tolerance mechanisms of stress at the molecular level. In our earlier study, a suppression subtractive hybridization (SSH) approach was used to isolate expressed sequence tags (ESTs) up-regulated during water-deficit stress in sugarcane (Prabu et al., 2010). Among the genes identified, *ScMYBAS1* (acc. no. EU670236) encoding a putative MYB transcription factor was found to be up-regulated by water- deficit and salinity stress (Prabu et al., 2010).

A large group of plant transcription factors such as AP2/EREBP, bZIP, WRKY, MYB and zinc finger proteins has been characterized as a part of gene regulation through which plants respond to biotic and abiotic stresses (Grotewold, 2008). Recently these transcription factors regulating stress responsive genes have become important target genes for improving plant stress tolerance. Among them, MYB family of genes/proteins constitute largest, functionally diverse group of plant transcription factor with varying numbers of MYB domain repeats conferring their ability to bind DNA. Members of this family are key factors in regulatory networks controlling development, metabolism and responses to biotic and abiotic stresses (Du et al., 2009). The elucidation of MYB protein structure - function relationship by computational or experimental methods will provide the foundation for predicting the contributions of MYB proteins to the biology of plants in general. Therefore, in this study, the three-dimensional structures of ScMYBAS1-3 was generated using the multiple threading alignments and iterative structural assembly simulations. Further, the predicted structural model was evaluated by using structure quality assessment programs.

2. Methodology

Three dimensional structure prediction

The iterative threading assembly refinement (I-TASSER) server (Roy et al., 2010) was used to predict the three dimensional structure of the ScMYBAS1-3. There were three stages involved in this procedure:

Stage 1: Threading

(i)The query protein sequence, ScMYBAS1-3 was matched against a nonredundant sequence database by position-specific iterated BLAST (PSI-BLAST; (Altschul et al., 1997)).

(ii)Based on multiple alignment of the sequence homologs, a sequence profile was created for prediction of secondary structure using PSIPRED (Jones, 1999).

(iii)Based on the sequence profile and the predicted secondary structure, ScMYBAS1-3 sequence then threaded through a representative PDB structure library using a locally installed meta-threading server (LOMETS) and the threading programs (FUGUE, HHSEARCH, MUSTER, PROSPECT, PPA, SP3 and SPARKS).

(iv)Based on the threading results, top template hit was then selected for further modelling. The quality of the template alignment was judged by statistically significant energy (Z-score).

Stage 2: Structural assembly: *Ab-initio* modeling

(i)Based on the threading alignments, aligned sections that aligned well was excised from the template structure.

(ii)The unaligned regions were modeled by *ab-initio* modeling (Jauch et al., 2007).

Stage 3: Model selection and refinement

(i)Fragment assembly simulation was performed again starting from the selected cluster centroids.

(ii)During this second round of simulations, the lowest energy structures were selected as input for REMO (Li and Zhang, 2009), which generates the final structural model through the optimization of hydrogen bonding networks.

Structure quality assessment

Based on the I-TASSER predictions, five different models of the ScMYBAS1-3

structure were generated. Finally one model was selected by the overall stereo chemical quality.

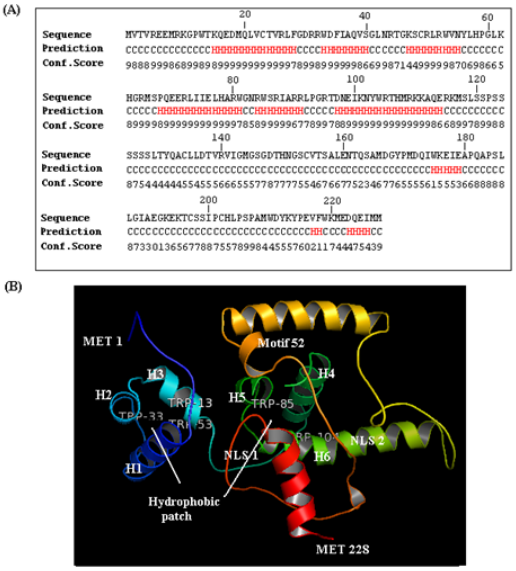
To test the quality of the model, the PROCHECK (Laskowski et al., 1993), Verify-3D (Eisenberg et al., 1997) and ProSA (Wiederstein and Sippl, 2007) programs were used. The stereo chemical quality of the five best scoring models were assessed by PROCHECK. The overall stereo chemical quality of the model was assessed by Ramachandran plot analysis. ProSA program (https://prosa.services.came.sbg.ac.at/prosa.php) was used to evaluate the energy of the structure using a distance-based pair potential.

3. Results and Discussion

Three dimensional structure of ScMYBAS1-3

In order to better characterize the ScMYBAS1-3 protein and unavailability of evolutionarily homologous template structure, threading and *ab initio* modelling was performed using I-TASSER platform (Roy et al., 2010). I-TASSER generates three-dimensional (3D) atomic models from multiple threading alignments and iterative structural assembly simulations. From the result of PSI-BLAST, a sequence profile was created based on multiple alignment of the sequence homologs. This sequence profile was used to predict the secondary structure of ScMYBAS1-3 with a statistically significant *Z* – score of > 1 using PSIPRED (**Fig. 1A**; Jones,1999).

Fig 1. Three dimensional structure of ScMYBAS1-3. (A) Predicted secondary structure of the ScMYBAS1-3. The prediction contains α-helix, coil structures are designated as H, C respectively with confidence scores for each residue (between 0 to 9) (B) predicted model of ScMYBAS1-3 using I-TASSER. The figure was produced using Pymol and Discovery studio visualizer 2.5. The α-helices in R2 and R3 repeats, hydrophobic patch, conserved tryptophan residues, nuclear localization signals (NLS1, NLS 2) and the transactivation motif (Motif 52) are labeled in white



In total five models were generated by I-TASSER server from which the model with lowest C-score (-3.60) was selected as the best (**Fig. 1B**) and subjected to internal evaluation of self-consistency checks. As described in I-TASSER (Roy et al., 2010), C-score is an estimate of the quality of the predicted model and confidence of structure prediction.

As a final test of the quality of the model (**Fig. 1**), the PROCHECK, Verify-3D, PROMOTIF and ProSA programs were used. The PROCHECK was applied to quantify the residues in available zones of the Ramachandran plot. Thus, 85.4 % of residues were located in the most favored zones, 10.1 % in allowed regions, 2 % (four residues)

in disallowed region. The Goodness factors (G-factors), from the PROCHECK results showed the quality of covalent and overall bond-angle distances. The comparable Ramachandran plot characteristic and the goodness factors confirm the quality of the modeled structure.

Verify-3D uses a score function to assess the quality of the model. The Verify-3D profile for the final structure showed that 86.4 % residues with a score over 0.2, thus the predicted ScMYBAS1-3 model was considered as reliable.

The identified structural motifs and their locations in the predicted model were evaluated using PROMOTIF program and presented in **Table 1**. Finally, ProSA program evaluates the energy of the structure using a distance-based pair potential. Residues with negative ProSA energies confirm the reliability of the predicted model. The *z*-score (-3.56) value is displayed in plot that contains the *z*-scores of all experimentally determined protein chains in current PDB and indicates overall model quality.

Table 1. Structural motifs identified in the predicted ScMYBAS1-3 structure using PROMOTIF program. (A) α-hleix (B) β- turn and (C) γ- turn motifs

(A)

Helix Number	Start Residue	End Residue	Helix Type	No. of Residues	Sequence
1	15	27	α	13	[illegible]
2	34	41	α	8	[illegible]
3	46	54	α	9	[illegible]
4	68	79	α	12	[illegible]
5	85	91	α	7	[illegible]
6	98	104	α	7	[illegible]
7	107	120	α	14	[illegible]
8	149	167	α	19	[illegible]
9	171	175	α	5	[illegible]
10	194	197	α	4	[illegible]
11	214	226	α	13	[illegible]

(B)

Residue numbers	Sequence	Turn Type	(i+1) Φ	(i+1) Ψ	(i+2) Φ	(i+2) Ψ	Φ Ψ Region	Chi1 (i+1)	Chi1 (i+2)	i to i+3 Distance
2 - 5	VTVR	I	-66.0	-37.1	-75.5	-30.6	AA	-167.1	64.8	5.4
5 - 8	REEM	IV	72.6	-0.4	-69.3	-20.4	LA	-71.5	-176.9	6.9
10 - 13	KGPW	IV	-66.9	-81.5	-83.6	140.5	aP	-	-27.8	6.7
29 - 32	GDRR	I	-56.1	-31.8	-65.2	-29.6	AA	-76.8	-161.9	5.8
30 - 33	DRRW	IV	-65.2	-29.6	-54.3	-54.6	AA	-161.9	178.3	5.9
41 - 44	GLNR	IV	82.8	-37.9	-140.8	25.8	Gα	-78.3	-169.1	3.3
54 - 57	VNYL	IV	-102.5	116.2	65.0	36.2	BL	54.5	-157.3	6.3
55 - 58	NYLH	IV	65.6	36.2	-113.9	-8.6	LA	-157.3	-73.3	6.9
79 - 82	ARWG	IV	-77.4	-15.8	-162.2	125.3	AB	-67.6	-162.2	5.9
81 - 84	WGNP	II	-85.2	126.3	52.3	25.3	PL	-	-83.9	5.7
92 - 95	LPGR	IV	-8.8	103.0	119.8	-4.5	G	-25.8	-	5.2
121 - 124	SPSS	IV	-61.7	-21.6	170.7	-37.2	A	33.8	-179.5	5.8
130 - 133	TYQA	IV	51.0	71.9	-9.7	90.2	L	-67.9	59.2	6.8
136 - 139	LDTV	IV	-60.9	-24.2	-54.6	-35.0	AA	-76.2	-58.4	5.7
137 - 140	DTVR	VIII	-54.6	-35.0	-137.4	119.3	AB	-58.4	-57.2	5.2
146 - 149	SGDT	IV	-93.7	-117.0	-58.6	141.5	BP	-	-169.5	6.5
167 - 170	DGYP	IV	-98.6	-108.5	-57.5	124.4	P	-	-174.6	6.4
175 - 178	WKEI	IV	72.3	30.5	47.3	89.6	GL	-65.6	60.6	6.0
176 - 179	KEIE	IV	47.3	89.6	51.1	59.3	LL	60.6	58.2	5.3
188 - 191	GIAE	IV	-61.1	-30.5	-56.8	-37.1	AA	-167.5	-	6.0
189 - 192	IAEG	IV	-56.8	-37.1	-59.6	-51.1	AA	-	-64.8	5.5
190 - 193	AEGK	IV	-59.6	-51.1	-71.9	95.0	AP	-64.8	-	5.7
204 - 207	LPSP	IV	-80.6	138.5	47.3	34.2	PL	16.7	-70.8	5.4
210 - 213	WDYK	IV	144.3	-63.7	-69.9	-30.7	A	-163.5	43.3	5.7
211 - 214	DYKY	I	-69.9	-30.7	-78.3	-19.6	AA	43.3	-68.2	5.2

(C)

Residue numbers	Sequence	Turn Type	(i+1) Φ	(i+1) Ψ	i to i+2 Distance
32 - 34	RWD	CLASSIC	57.2	-57.4	6.0
96 - 98	TDN	CLASSIC	80.6	-47.7	5.6
143 - 145	GMG	INVERSE	-76.5	92.3	5.5
169 - 171	VPM	INVERSE	-62.9	80.7	5.5
191 - 193	EGK	INVERSE	-71.9	95.0	5.2

Overall architecture of ScMYBAS1-3 model

The ScMYBAS1-3 structure contains 11 α-helices, 25 β-turns and five γ- turn motifs (**Table. 1A, B and C**). The ScMYBAS1-3 contains two imperfect tandem repeats of 53 and 49 amino acids, designated as R2 and R3. As shown in **Table 2**, R2 and R3 repeat regions consists of three well defined α-helices and two less-well defined turns respectively forming a helix-turn-helix (HLH) structure (**Fig. 1B**). Three and two regularly spaced (20 residues apart) tryptophan residues were present in R2 and R3 repeat respectively (**Fig. 1B**), which forms a tryptophan cluster in three-dimensional HLH structure (**Table 2; Fig. 1B**) and critical in sequence-specific binding. These structures are in accordance with the reported characteristics of typical plant R2R3 MYB proteins (Ogata et al., 1992; Ying et al., 2000). The third α-helices predicted in R2 and R3 repeats make contact with the DNA bases and thus determine the binding specificity (Ogata et al., 1992; 1994).

Table 2. Overall structural architecture of the predicted ScMYBAS1-3 model as shown in Fig. 1B

Structural component	Type of structural motif	Residue Locations
R2 repeat	α – Helix	Lys^{15} to Leu^{27}
	β-Turn – Type IV	Asp^{30} to Trp^{33}
	α – Helix	Asp^{34} to Gly^{41}
	β-Turn - Type IV	Leu^{42} to Thr^{45}
	α – Helix	Gly^{46} to Val^{54}
R3 repeat	α – Helix	Pro^{68} to Arg^{80}
	β-Turn – Type II	Trp^{81} to Arg^{84}
	α – Helix	Trp^{85} to Arg^{91}
	β-Turn – Type IV	Leu^{92} to Arg^{95}
	α – Helix	Asn^{98} to Trp^{104}
Nuclear localization signal (NLS1)	β-Turn – Type IV	Lys^{62} to Arg^{65}
Nuclear localization signal (NLS2)	α – Helix	Arg^{109} to Met^{117}
Trans-activator motif	α – Helix	Met^{171} to Trp^{175}

4. Conclusion

Present computational modelling study, has provided structure of DNA binding sugarcane transcription factor (ScMYBAS1), which would definitely assist structure based docking studies to accelerate the search of interacting DNA regions and or proteins.

Acknowledgement

The authors are thankful to Director General, VSI, India for encouragement and support during the course of study. Financial assistance from VSI, India is also gratefully acknowledged and Dr. S. Anandhan, Scientist 'C', DARL, Defense Research and Developmental Organization, Haldwani, India for his valuable suggestions during the research work.

References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25: 3389-3402.

Du, H., Zhang, L., Liu, L., Tang, X.F., Yang, W.J., Wu, Y.M., Huang, Y.B., Tang, Y.X. 2009. Biochemical and molecular characterization of plant MYB transcription factor family. Biochem (Moscow), 74: 1-11.

Eisenberg, D., Lüthy, R., Bowie, J.U. 1997. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol, 277: 396-404.

Grotewold, E. 2008. Transcription factors for predictive plant metabolic engineering: are we there yet?. Curr Opin Biotechnol, 19: 138-144.

Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D. 2007. Assessment of CASP7 structure predictions for template free targets. Proteins, 69: 57-67.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, 292: 195-202.

Kryshtafovych, A., Fidelis, K. 2009. Protein structure prediction and model quality assessment. Drug Discov Today, 14: 386-393.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst, 26: 283-291.

Li, Y., Zhang, Y. 2009. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins, 76: 665-676.

Menossi, M., Silva-Filho, M.C., Vincentz, M., Van-Sluys, M.A., Souza, G.M. 2008. Sugarcane functional genomics: gene discovery for agronomic trait development. Int J Plant Genomics, 2008: 458732.

Murzin, A.G. 2001. Progress in protein structure prediction. Nat Struct Biol, 8:110-112.

Ogata, K., Hojo, H., Aimoto, S., Nakai, T., Nakamura, H., Sarai, A., Ishii, S., Nishimura, Y. 1992. Solution structure of

a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. Proc Natl Acad Sci USA, 89: 6428-6432.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S., Nishimura, Y. 1994. Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. Cell, 79: 639-648.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M.S., Davis, F.P., Stuart, A.C., Mirkovic, N., Rossi, A., Marti-Renom, M.A., Fiser, A., Webb, B., Greenblatt, D., Huang, C.C., Ferrin, T.E., Sali, A. 2004. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res, 32: D217-222.

Pierri, C.L., Parisi, G., Porcelli, V. 2010. Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening. Biochim Biophys Acta, 1804:1695-1712.

Prabu, G.R., Kawar, P.G., Madhuri, C.P., Theertha Prasad, D. 2010. Identification of water deficit stress upregulated genes in sugarcane. Plant Mol Biol Rep, (DOI 10.1007/s11105-010-0230-0).

Roy, A., Kucukural, A., Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc, 5: 725-738.

Wiederstein, M., Sippl M.J. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res, 35 (Web Server issue): W407-410.

Ying, G.G., Proost, P., van Damme, J., Bruschi, M., Introna, M., Golay, J. 2000. Nucleolin, a novel partner for the Myb transcription factor family that regulates their activity. J Biol Chem, 275: 4152-4158.

Zhang, Y. 2009. Protein structure prediction: when is it useful? Curr Opin Struct Biol, 19:145-155.